Object Detection and tracking in Video Sequences

Manisha Chate[1] , S.Amudha[2] ,Vinaya Gohokar[3]
[1]Amity University, Noida ,India
manisha_157@rediffmail.com
[2]Amity University, Noida, India
amudhaece@gmail.com
[3]SSGMCE, Shegaon , India
vvgohokar@rediffmail.com

***Abstract*— This paper focuses on key steps in video analysis i.e. Detection of moving objects of interest and tracking of such objects from frame to frame. The object shape representations commonly employed for tracking are first reviewed and the criterion of feature Selection for tracking is discussed. Various object detection and tracking approaches are compared and analyzed.**

***Index Terms*—feature selection, image segmentation, object representation ,point tracking**

I. Introduction

Videos are actually sequences of images, each of which called a frame, displayed in fast enough frequency so that human eyes can percept the continuity of its content. It is obvious that all image processing techniques can be applied to individual frames. Besides, the contents of two consecutive frames are usually closely related.

Visual content can be modeled as a hierarchy of abstractions. At the first level are the raw pixels with color or brightness information. Further processing yields features such as edges, corners, lines, curves, and color regions. A higher abstraction layer may combine and interpret these features as objects and their attributes. At the highest level are the human level concepts involving one or more objects and relationships among them

Object detection in videos involves verifying the presence of an object in image sequences and possibly locating it precisely for recognition. Object tracking is to monitor objects spatial and temporal changes during a video sequence, including its presence, position, size, shape, etc.

This is done by solving the temporal correspondence problem, the problem of matching the target region in successive frames of a sequence of images taken at closely-spaced time intervals. These two processes are closely related because tracking usually starts with detecting objects, while detecting an object repeatedly in subsequent image sequence is often necessary to help and verify tracking.

II. Object Detection and Tracking Approaches

A. Object Representation

In a tracking scenario, an object can be defined as anything that is of interest for further analysis. For instance, boats on the sea, fish inside an aquarium, vehicles on a road, planes in the air, people walking on a road, or bubbles in the water are a set of objects that may be important to track in a specific domain. Objects can be represented by their shapes and appearances. In this section, we will first describe the object shape representations commonly employed for tracking and then address the joint shape and appearance representations.

—*Points.* The object is represented by a point, that is, the centroid (Figure 1(a))[2] or by a set of points (Figure 1(b)) [3].

In general, the point representation is suitable for tracking objects that occupy small regions in an image.

—*Primitive geometric shapes.* Object shape is represented by a rectangle, ellipse (Figure 1(c), (d) [4]. Object motion for such representations is usually modeled by translation, affine, or projective (homography) transformation. Though primitive geometric shapes are more suitable for representing simple rigid objects, they are also used for tracking non rigid objects.

—*Object silhouette and contour.* Contour representation defines the boundary of an object (Figure 1(g), (h). The region inside the contour is called the silhouette of the object (see Figure 1(i)). Silhouette and contour representations are suitable for tracking complex no rigid shapes [5].

—*Articulated shape models.* Articulated objects are composed of body parts that are held together with joints. For example, the human body is an articulated object with torso, legs, hands, head, and feet connected by joints. The relationship between the parts is governed by kinematic motion models, for example, joint angle, etc. In order to represent an articulated object, one can model the constituent parts using cylinders or ellipses as shown in Figure 1(e).

—*Skeletal models.* Object skeleton can be extracted by applying medial axis transform to the object silhouette [6]. This model is commonly used as a shape representation for recognizing objects [7]. Skeleton representation can be used to model both articulated and rigid objects (see Figure 1(f).

DOI: 01.IJSIP.03.01.84

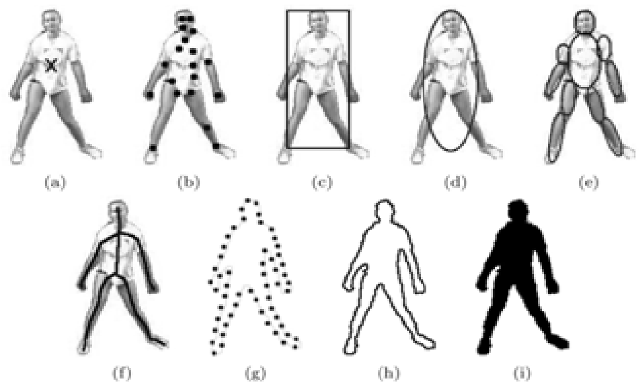


Fig 1. Object representations. (a) Centroid, (b) multiple points, (c) rectangular patch, (d) elliptical patch, (e) part-based multiple patches, (f) object skeleton, (g)complete object contour, (h) control points on object contour, (i) object silhouette.

B. Feature Selection For Tracking

Selecting the right features plays a critical role in tracking. In general, the most desirable property of a visual feature is its uniqueness so that the objects can be easily distinguished in the feature space. Feature selection is closely related to the object representation.For example, color is used as a feature for histogram-based appearance representations, while for contour-based representation, object edges are usually used as features. In general, many tracking algorithms use a combination of these features. The details of common visual features are as follows.

—*Color.* The apparent color of an object is influenced primarily by two physical factors,
1) the spectral power distribution of the illuminant and 2) the surface reflectance properties of the object. In image processing, the RGB (red, green, blue) color space is usually used to represent color. However, the RGB space is not a perceptually uniform color space, that is, the differences between the colors in the RGB space do not correspond to the color differences perceived by humans [8]. Additionally, the RGB dimensions are highly correlated. In contrast, *L*”*u*”*v*” and *L*”*a*”*b*” are perceptually uniform color paces, while HSV (Hue, Saturation, Value) is an approximately uniform color space However, these color spaces are sensitive to noise [9]. In summary, there is no last word on which color space is more efficient, therefore a variety of color spaces have been used in tracking.

—*Edges.* Object boundaries usually generate strong changes in image intensities. Edge
detection is used to identify these changes. An important property of edges is that they are less sensitive to illumination changes compared to color features. Algorithms that track the boundary of the objects usually use edges as the representative feature. Because of its simplicity and accuracy, the most popular edge detection approach is the Canny Edge detector [10]. An evaluation of the edge detection algorithms is provided by [11].

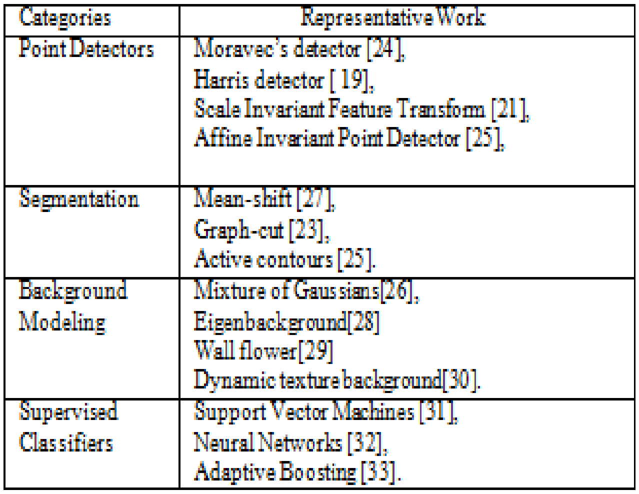
—*Optical Flow.* Optical flow is a dense field of displacement vectors which defines the translation of each pixel in a region. It is computed using the brightness constraint, which assumes brightness constancy of corresponding pixels in consecutive frames [12]. Optical flow is commonly used as a feature in motion-based segmentation and tracking applications.

—*Texture.* Texture is a measure of the intensity variation of a surface which quantifies properties such as smoothness and regularity. Compared to color, texture requires a processing step to generate the descriptors. There are various texture descriptors:

Gray-Level Co occurrence Matrices (GLCM’s) [13] (a 2D histogram which shows the co occurrences of intensities in a specified direction and distance),Law’s texture measures [14] (twenty-five 2D filters generated from five 1D filters corresponding to level, edge, spot, wave, and ripple), wavelets [15] (orthogonal bank of filters), and steerable pyramids [16]. Similar to edge features, the texture features are less sensitive to illumination changes compared to color..Mostly features are chosen manually by the user depending on the application domain. However, the problem of automatic feature selection has received significant attention in the pattern recognition community. Automatic feature selection methods can be divided into *filter* methods and *wrapper* methods [17]. The filter methods try to select the features based on a general criteria, for example, the features should be uncorrelated. The wrapper methods select the features based on the usefulness of the features in a specific problem domain, for example, the classification performance using a subset of features.

Among all features, color is one of the most widely used feature for tracking. Despite its popularity, most color bands are sensitive to illumination variation. Hence in scenarios where this effect is inevitable, other features are incorporated to model object appearance. Alternatively, a combination of these features is also utilized to improve the tracking performance

Table I. Object Detection Categories



Categories	Representative Work
Point Detectors	Moravec’s detector [24], Harris detector [19], Scale Invariant Feature Transform [21], Affine Invariant Point Detector [25],
Segmentation	Mean-shift [27], Graph-cut [23], Active contours [25].
Background Modeling	Mixture of Gaussians[26], Eigenbackground[28] Wall flower[29] Dynamic texture background[30].
Supervised Classifiers	Support Vector Machines [31], Neural Networks [32], Adaptive Boosting [33].

III. OBJECT DETECTION

Every tracking method requires an object detection mechanism either in every frame or when the object first appears in the video. A common approach for object detection is to use information in a single frame. However, some object detection methods make use of the temporal information computed from a sequence of frames to reduce the number of false detections. This temporal information is usually in the form of frame differencing, which highlights changing regions in consecutive frames. Given the object regions in the image, it is then the tracker's task to perform object correspondence from one frame to the next to generate the tracks.

A. Point Detectors

Point detectors are used to find interest points in images which have an expressive texture in their respective localities. Interest points have been long used in the context of motion, stereo, and tracking problems. A desirable quality of an interest point is its invariance to changes in illumination and camera viewpoint. In the literature, commonly used interest point detectors include Moravec's interest operator [18], Harris interest point detector [19], KLT detector [20], and SIFT detector [21] as illustrated in figure 2.

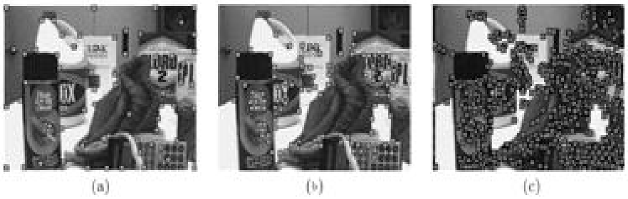


Fig 2. Interest points detected by applying (a) the Harris, (b) the KLT, and (c) SIFT operators

B. Background Subtraction

Object detection can be achieved by building a representation of the scene called the background model and then finding deviations from the model for each incoming frame. Any significant change in an image region from the background model signifies a moving object. The pixels constituting the regions undergoing change are marked for further processing. Usually, a connected component algorithm is applied to obtain connected regions corresponding to the objects. This process is referred to as the *background subtraction*.

For instance, Stauffer and Grimson [21] use a mixture of Gaussians to model the pixel color. In this method, a pixel in the current frame is checked against the background model by comparing it with every Gaussian in the model until a matching Gaussian is found. If a match is found, the mean and variance of the matched Gaussian is updated, otherwise a new Gaussian with the mean equal to the current pixel color and some initial variance is introduced into the mixture. Each pixel is classified based on whether the matched distribution represents the background process. Moving regions, which are detected using this approach, along with the background models are shown in Figure 3.

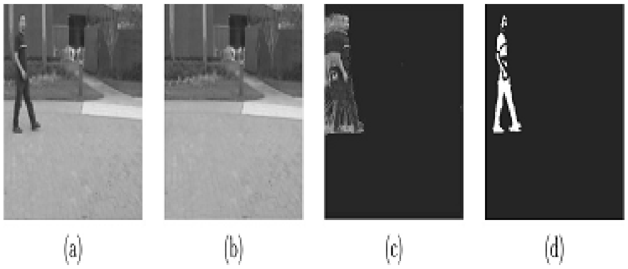


Fig 3. Mixture of Gaussian modeling for background subtraction. (a) Image from a sequence in which a person is walking across the scene. (b) The mean of the highest-weighted Gaussians at eachpixels position. These means represent themosttemporallypersistent per-pixel color and hence shouldrepresentthe stationary background. (c) The means of the Gaussian with the second-highest weight; these means represent colors that are observed less frequently. (d) Background subtractionresult. The foreground consists of the pixels in the current frame that matched a low-weighted Gaussian.

Another approach is to incorporate region-based (spatial) scene information instead of only using color-based information. Elgammal and Davis [22] use nonparametric kernel density estimation to model the per-pixel background. During the subtraction process, the current pixel is matched not only to the corresponding pixel in the background model, but also to the nearby pixel locations. Thus, this method can handle camera jitter or small movements in the background. Li and Leung [2002] fuse the texture and color features to perform background subtraction over blocks of 5 × 5 pixels. Since texture does not vary greatly with illumination changes, the method is less sensitive to illumination. Toyama et al. [1999] propose a three-tiered algorithms to deal with the background subtraction problem. In addition to the pixel-level subtraction, the authors use the region and the frame-level information. At the pixel level, the authors propose to use Wiener filtering to make probabilistic predictions of the expected background color. At the region level, foreground regions consisting of homogeneous color are filled in. At the frame level, if most of the pixels in a frame exhibit suddenly change, it is assumed that the pixel-based color background models are no longer valid. At this point, either a previously stored pixel-based background model is swapped in, or the model is reinitialized. The foreground objects are detected by projecting the current image to the eigenspace and finding the difference between the reconstructed and actual images. We show detected object regions using the eigenspace approach in Figure 4.

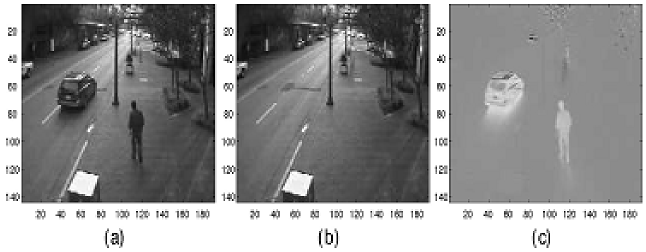


Figure 4. Eigenspace decomposition-based background ubtraction (space is constructed with objects in the FOV of camera): (a) an input image with objects, (b) reconstructed image after projecting input image onto the eigenspace, (c) difference image. Note that the foreground objects are clearly identifiable.

DOI: 01.IJSIP.03.01. 84

C. Segmentation

The aim of image segmentation algorithms is to partition the image into perceptually similar regions. Every segmentation algorithm addresses two problems, the criteria for a good partition and the method for achieving efficient partitioning [23].

1. Mean-Shift Clustering.

For the image segmentation problem, Comaniciu and Meer [2002] propose the mean-shift approach to find clusters in the joint spatial color space, [*l* , *u*, *v*, *x*, *y*], where [*l* , *u*, *v*] represents the color and [*x*, *y*] represents the spatial location. Given an image, the algorithm is initialized with a large number of hypothesized cluster centers randomly chosen from the data. Then, each cluster center is moved to the mean of the data lying inside the multidimensional ellipsoid centered on the cluster center. The vector defined by the old and the new cluster centers is called the *mean-shift vector*. The mean-shift vector is computed iteratively until the cluster centers do not change their positions. Note that during the mean-shift iterations, some clusters may get merged. In Figure 5(b), we show the segmentation using the mean-shift approach generated using the source code available at Mean Shift Segments

2. Image Segmentation Using Graph-Cuts.

Image segmentation can also be formulated as a graph partitioning problem, where the vertices (pixels), **V** = {*u*, *v*, . . .}, of a graph (image), **G**, are partitioned into *N* disjoint sub-graphs (regions),

$$Ai \ , \ \_N \ i = 1 \ Ai = \mathbf{V}, \ Ai \cap Aj = \varnothing \ , \ i \ \_= j \ ,$$

by pruning the weighted edges of the graph. The total weight of the pruned edges between two sub graphs is called a *cut*. The weight is typically computed by color, brightness, or texture similarity between the nodes. Wu and Leahy [1993] use the minimum cut criterion, where the goal is to find the partitions that minimize a cut.In their approach, the weights are defined based on the color similarity. One limitation of minimum cut is its bias toward over segmenting the image. This effect is due to the increase in cost of a cut with the number of edges going across the two partitioned segments.

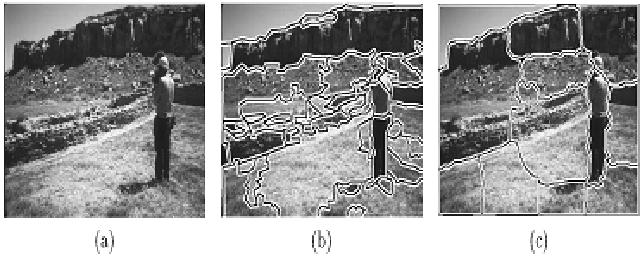


Fig 5. Segmentation of the image shown in (a), using mean-shift segmentation (b) and normalized cuts (c)

IV. OBJECT TRACKING

The aim of an object tracker is to generate the trajectory of an object over time by locating its position in every frame of the video. Object tracker may also provide the complete region in the image that is occupied by the object at every time instant. The tasks of detecting the object and establishing correspondence between the object instances across frames can either be performed separately or jointly. In the first case, possible object regions in every frame are obtained by means of an object detection algorithm, and then the tracker corresponds objects across frames. In the latter case, the object region and correspondence is jointly estimated by iteratively updating object location and region information obtained from previous frames. In either tracking approach, the objects are represented using the shape and/or appearance models described in Section 2. The model selected to represent object shape limits the type of motion or deformation it can undergo. For example, if an object is represented as a point, then only a translational model can be used. In the case where a geometric shape representation like an ellipse is used for the object, parametric motion models like affine or projective transformations are appropriate.

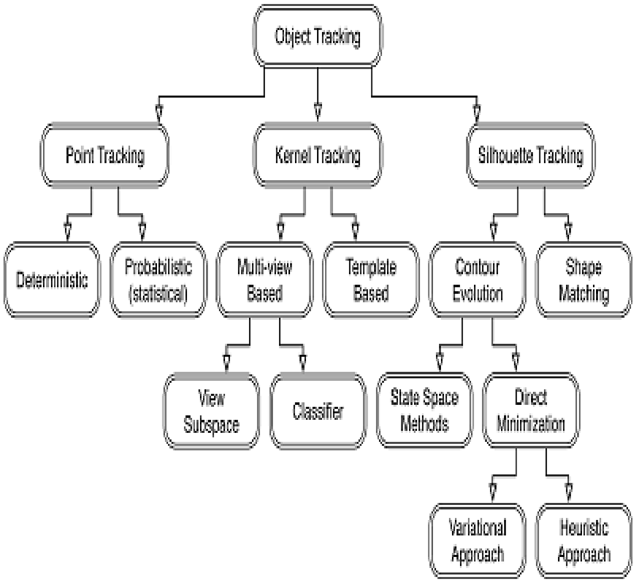


Fig 6. Taxonomy of tracking method.

In view of the aforementioned discussion, we provide taxonomy of tracking methods in Figure 6. Representative work for each category is tabulated in Table II. We now briefly introduce the main tracking categories, followed by a detailed section on each category.

Table. II

Categories	Representative Work
Point Tracking	
Deterministic methods Statistical methods	MGE tracker[35],GOA tracker [34]. Kalman filter[36], PMHT[36].
Kernel Tracking	
Template and density based appearance models	Mean-shift[4],KLT[20], Layering [37].
Multi –view appearance models	Eigentracking [lack and jepon 1998]; SVM tracker [38].
Silhouete Traxking	
Contour evolution	State space models[39], Variatinal methods[40], Heuristic methods[41].
Matching shapes	Hausdorff[42] Histogram[43].

—Point Tracking. Objects detected in consecutive frames are represented by points,and the association of the points is based on the previous object state which can include object position and motion. This approach requires an external mechanism to detect the objects in every frame. An example of object correspondence is shown inFigure 7(a).

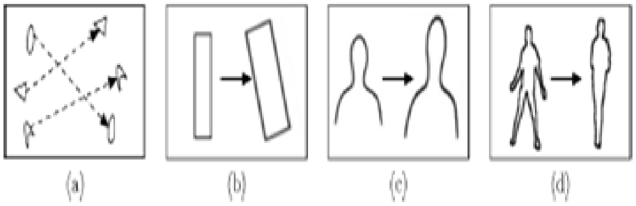


Fig 7. (a) Different tracking approaches. Multipoint correspondence, (b) parametric transformation of a rectangular patch, (c, d) Two examples of contour evolution.

—Kernel Tracking. Kernel refers to the object shape and appearance. For example, the kernel can be a rectangular template or an elliptical shape with an associated histogram. Objects are tracked by computing the motion of the kernel in consecutive frames (Figure 7(b)). This motion is usually in the form of a parametric transformation such as translation, rotation, and affine.

—Silhouette Tracking. Tracking is performed by estimating the object region in each frame. Silhouette tracking methods use the information encoded inside the object region. This information can be in the form of appearance density and shape models which are usually in the form of edge maps. Given the object models, silhouettes are tracked by either shape matching or contour evolution (see Figure 7(c), (d)). Both of these methods can essentially be considered as object segmentation applied in the temporal domain using the priors generated from the previous frames.

Conclusion

An extensive survey of object detection and tracking methods is presented in this paper. Recognizing the importance of object shape representations for detection and tracking systems, we have included discussion on popular methods for the same. A detailed summary of criteria for feature selection ,object tracking methods is presented which can give valuable insight into this important research topic.

References

[1] GREGORY D. HAGER and Peter N. Belhumeur,Efficient Region Tracking With Parametric Models of Geometry and Illumination.IEEE transactions on pattern analysis and machine intelligence, vol. 20, no. 10, pp. 1025-39 October 1998.

[2] VEENMAN, C., REINDERS, M., AND BACKER, E. 2001. Resolving motion correspondence for densely moving points. *IEEE Trans. Patt. Analy. Mach. Intell. 23*, 1, 54–72.

[3] SERBY, D., KOLLER-MEIER, S., AND GOOL, L. V. 2004. Probabilistic object tracking using multiple features. In *IEEE International Conference of Pattern Recognition (ICPR)*. 184–187.

[4] CHUNHUA SHEN,JUNAE KIM AND HANZI WANG, Generalised Kernel –Based visual Tracking .IEEE transaction on circuit and system for video technology,Vol.20,no.1,January 2010

[5] YILMAZ, A., LI, X., AND SHAH, M. 2004. Contour based object tracking with occlusion handling in video acquired using mobile cameras. *IEEE Trans. Patt. Analy. Mach. Intell. 26*, 11, 1531–1536.

[6] BALLARD, D. AND BROWN, C. 1982. *Computer Vision*. Prentice-Hall.

[7] ALI, A. AND AGGARWAL, J. 2001. Segmentation and recognition of continuous human activity. In *IEEEWorkshop on Detection and Recognition of Events in Video*. 28–35.

[8] PASCHOS, G. 2001. Perceptually uniform color spaces for color texture analysis: an empirical evaluation. *IEEE Trans. Image Process. 10*, 932–937.

[9] SONG, K. Y., KITTLER, J., AND PETROU, M. 1996. Defect detection in random color textures. *Israel Verj. Cap.J. 14*, 9, 667–683

[10] CANNY, J. 1986. A computational approach to edge detection. *IEEE Trans. Patt. Analy. Mach. Intell. 8*, 6,679–698

[11] BOWYER, K., KRANENBURG, C., AND DOUGHERTY, S. 2001. Edge detector evaluation using empirical roc curve.*Comput. Vision Image Understand. 10*, 77–103

[12] SYNH VIET-UYEN HAAND JAE WOOK JEON .Readjusting Unstable Regions to Improve the Quality of High Accuracy Optical Flow. IEEE transaction on circuit and system for video technology, Vol. 20, NO. 4, APRIL 2010

[13] HARALICK, R., SHANMUGAM, B., AND DINSTEIN, I. 1973. Textural features for image classification. *IEEE Trans.Syst. Man Cybern. 33*, 3, 610–622

[14] LAWS, K. 1980. Textured image segmentation. PhD thesis, Electrical Engineering, University of Southern California

[15] MALLAT, S. 1989. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Patt. Analy. Mach. Intell. 11*, 7, 674–693.

[16] GREENSPAN, H., BELONGIE, S., GOODMAN, R., PERONA, P., RAKSHIT, S., AND ANDERSON, C. 1994. Overcomplete steerable pyramid filters and rotation invariance. In *IEEE Conference on Computer Vision and PatternRecognition (CVPR)*. 222–228.

[17] BLUM, A. L. AND LANGLEY, P. 1997. Selection of relevant features and examples in machine learning. *Artific.Intell. 97*, 1-2, 245–271

[18] MORAVEC, H. 1979. Visual mapping by a robot rover. In *Proceedings of the International Joint Conferenceon Artificial Intelligence (IJCAI)*. 598–600

[19] HARRIS, C. AND STEPHENS, M. 1988. A combined corner and edge detector. In *4th Alvey Vision Conference*.147–151.

[20] SHI, J. AND TOMASI, C. 1994. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 593–600.

[21] LOWE, D. 2004. Distinctive image features from scale-invariant key points. *Int. J. Comput. Vision 60*, 2,91–110.

[22] ELGAMMAL, A., HARWOOD, D., AND DAVIS, L. 2000. Non-parametric model for background subtraction. In *European Conference on Computer Vision (ECCV)*. 751–767.

[23] SHI, J. AND MALIK, J. 2000. Normalized cuts and image segmentation. *IEEE Trans. Patt. Analy. Mach. Intell. 22*, 8, 888–905.

[24] MORAVEC, H. 1979. Visual mapping by a robot rover. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 598–600

[25] CASELLES, V., KIMMEL, R., AND SAPIRO, G. 1995. Geodesic active contours. In *IEEE International Conference on Computer Vision (ICCV)*. 694–699.

[26] STAUFFER, C. AND GRIMSON, W. 2000. Learning patterns

DOI: 01.IJSIP.03.01. 84

of activity using real time tracking. *IEEE Trans. Patt. Analy. Mach. Intell. 22*, 8, 747–767.

[27] COMANICIU, D. AND MEER, P. 1999. Mean shift analysis and applications. In *IEEE International Conferenceon Computer Vision (ICCV)*. Vol. 2. 1197–1203.

[28] OLIVER, N., ROSARIO, B., AND PENTLAND, A. 2000. A bayesian computer vision system for modeling human interactions. *IEEE Trans. Patt. Analy. Mach. Intell. 22*, 8, 831–843

[29] TOYAMA, K., J. KRUMM, B. B., AND MEYERS, B. 1999. Wallflower: Principles and practices of background maintenance. In *IEEE International Conference on omputer Vision (ICCV)*. 255–261.

[30] MONNET, A., MITTAL, A., PARAGIOS, N., AND RAMESH, V. 2003. Background modeling and subtraction of dynamic scenes. In *IEEE International Conference on Computer Vision (ICCV)*. 1305–1312.

[31] PAPAGEORGIOU, C., OREN, M., AND POGGIO, T. 1998. A general framework for object detection. In *IEEE International Conference on Computer Vision (ICCV)*. 555–562.

[32] ROWLEY, H., BALUJA, S., ANDKANADE, T. 1998. Neural network-based face detection. *IEEE Trans. Patt. Analy.Mach. Intell. 20*, 1, 23–38.

[33] VIOLA, P., JONES, M., AND SNOW, D. 2003. Detecting pedestrians using patterns of motion and appearance.In *IEEE International Conference on Computer Vision (ICCV)*. 734–741.

[34] SALARI, V. AND SETHI, I. K. 1990. Feature point correspondence in the presence of occlusion. *IEEE Trans.Patt. Analy. Mach. Intell. 12*, 1, 87–91.

[35] BROIDA, T. AND CHELLAPPA, R. 1986. Estimation of object motion parameters from noisy images. *IEEE Trans.Patt. Analy. Mach. Intell. 8*, 1, 90–99.

[36] STREIT, R. L. AND LUGINBUHL, T. E. 1994. Maximum likelihood method for probabilistic multi-hypothesis tracking. In *Proceedings of the International Society for Optical Engineering (SPIE.)* vol. 2235. 394–405.

[37] TAO, H., SAWHNEY, H., AND KUMAR, R. 2002. Object tracking with bayesian estimation of dynamic layer representations. *IEEE Trans. Patt. Analy. Mach. Intell. 24*, 1, 75–89.

[38] AVIDAN, S. 2001. Support vector tracking. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 184–191.

[39] ISARD, M. AND BLAKE, A. 1998. Condensation - conditional density propagation for visual tracking. Int. J.Comput. Vision 29, 1, 5–28.

[40] BERTALMIO, M., SAPIRO, G., AND RANDALL, G. 2000. Morphing active contours. *IEEE Trans. Patt. Analy. Mach.Intell. 22*, 7, 733–737.

[41] RONFARD, R. 1994. Region based strategies for active contour models. *Int. J. Comput. Vision 13*, 2, 229–251.

[42] HUTTENLOCHER, D., NOH, J., AND RUCKLIDGE, W. 1993. Tracking nonrigid objects in complex scenes. In *IEEE International Conference on Computer Vision (ICCV)*. 93–101.

[43] KANG, J., COHEN, I., AND MEDIONI, G. 2004. Object reacquisition using geometric invariant appearance model. In *International Conference on Pattern Recongnition (ICPR)*. 759–762.

DOI: 01.IJSIP.03.01.84